C-Slate: A Multi-Touch and Object Recognition System for Remote Collaboration using Horizontal Surfaces

Shahram Izadi, Ankur Agarwal, Antonio Criminisi,
John Winn, Andrew Blake, Andrew Fitzgibbon
Microsoft Research Cambridge, 7 JJ Thomson Avenue, Cambridge, CB3 0FB
{shahrami, ankagar}@microsoft.com

Abstract

We introduce C-Slate, a new vision-based system, which utilizes stereo cameras above a commercially available tablet technology to support remote collaboration. The horizontally mounted tablet provides the user with high resolution stylus input, which is augmented by multi-touch interaction and recognition of untagged everyday physical objects using new stereo vision and machine learning techniques. This provides a novel and interesting interactive tabletop arrangement, capable of supporting a variety of fluid multi-touch interactions, including symmetric and asymmetric bimanual input, coupled with the potential for incorporating tangible objects into the user interface. When used in a remote context, these features are combined with the ability to see visual representations of remote users' hands and remote physical objects placed on top of the surface. This combination of bimanual and tangible interaction and sharing of remote gestures and physical objects provides a new way to collaborate remotely, complementing existing channels such as audio and video conferencing.

1. Introduction

Interactive tabletops can very naturally support co-located face-to-face collaboration [24, 25, 26]. This is a product of both their physical form, allowing users to view a large horizontal workspace whilst maintaining awareness of others, and their ability to support more fluid and direct interaction with digital content using touch, hands, gestures, and potentially other sensed physical objects [21, 23, 37]. Although the use of tabletops for co-located collaboration is well established, the role that such systems can play in remote collaboration has only very recently become a theme of investigation [6, 10, 32, 38].

Rich physical interactions occur over and around the tabletop surface during co-located collaboration. For example, we use our hands to refer to virtual or physical artifacts on the surface and use gestures to express actions. We constantly use the physical affordances of the tabletop [30], placing artifacts such as documents onto the surface to share and refer to them.

These physical interactions form important visual cues for collaboration, providing awareness of other peoples' actions and intentions, and facilitating fine grained coordination amongst members of the group. These interactions are clearly lost when we move to the remote case, and are often overlooked by current CSCW and groupware tools such as video and audio conferencing or remote desktop systems.

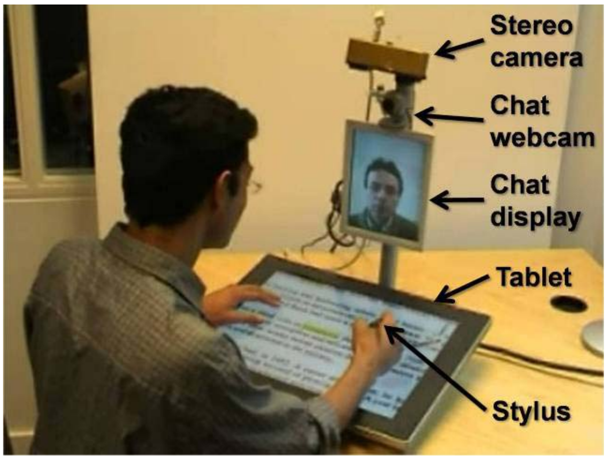

Figure 1: The C-Slate hardware setup comprising of a top-down stereo camera attached to a large tablet workspace, with a video conferencing display at eye-level to the user.

We introduce the Collaborative Slate (or C-Slate), a new vision-based system which utilizes a stereo camera attached above a commercially available horizontally mounted tablet technology as shown in Figure 1; with a second smaller display and web camera attached at eye-level to the user, supporting video and audio conferencing. The tablet supports high resolution stylus input, for fine grained inking and cursor control. Images captured from the stereo camera are processed using new vision and machine learning algorithms to allow high precision fingertip and touch detection, enabling both stylus and multi-touch interaction on the tablet surface.

New stereo vision and machine learning techniques are also used to recognize a variety of physical objects placed on the display, without the need to tag these with visual markers. This allows ordinary everyday objects such as post-its, documents, stationery or mobile devices to be sensed in order to invoke particular UI actions. This recognition system is also capable of detecting a user's hand poses, allowing these to also be mapped onto UI actions. These features provide a new and interesting interactive tabletop configuration, enabling a variety of fluid multi-touch interactions, including symmetric and asymmetric bimanual input, coupled with the potential for incorporating tangible objects into the UI.

The use of a top-down camera also allows images of hands and objects on top of the tablet surface to be rapidly captured, segmented and transmitted over the network to other C-Slates where they are rendered on the tablet display. This provides a virtual embodiment of the remote participant's hands, and allows remote parties to share images of physical objects with each other, such as written notes, drawings or game pieces, simply by placing them on the surface. This combination of bimanual and tangible interaction and sharing of remote gestures and physical objects provides a new way to collaborate remotely, complementing already established channels such as audio and video conferencing.

2. Related work

Our system relates to a large body of work within multi-touch and object sensing, direct input and tangible tabletops, and remote gesturing tools. We shall cover these aspects in turn within this section.

Multi-touch has received a great deal of attention recently through the widely disseminated research of Wilson [36, 37] and Han [13], and products such as the Apple iPhone [2] and now the Microsoft Surface [21]. Multi-touch has an even longer history however, and the first systems appeared well over two decades ago (see [5] for an overview of the major landmarks).

One technique of detecting multiple fingertips on a display is to build custom sensing electronics into the surface itself [8, 18, 23, 35]. These systems are typically based on capacitive sensing, although other sensors can be utilized [18, 27]. They usually sense at low resolutions and are visually opaque, relying on projection for display. Even with this low-resolution sensing, rich sets of interactions have been demonstrated [23, 40]. What is harder with such systems (as they are non-optical) is to image the entire hand or other arbitrary physical objects close to or touching the surface. Rather, other objects aside from the hand need to be actively tagged to be detected by the surface [9, 23].

Camera-based systems allow more flexibility in sensing, providing a higher resolution optical system for capturing richer information about arbitrary objects in proximity to the display. Wilson [37] clearly highlights the tradeoffs of this flexibility, in terms of the high computational costs, the difficulty in achieving real-time interactive rates, ambiguity of data (particularly detecting when an object is hovering as opposed to touching the surface), and susceptibility to occlusion and adverse lighting conditions. This makes developing such systems an interesting and challenging problem.

2.1 Camera based Multi-touch Systems

One common approach to building a multi-touch and object sensing tabletop is to place a camera on top or underneath the display surface, and use computer vision algorithms to process the captured images. These top-down and bottom-up configurations carry various tradeoffs. Top-down approaches [20, 34, 37, 39] tend to capture richer data regarding the interactions occurring on the surface as the camera is directly pointing at the display, although occlusion can be an issue. The camera can feasibly image and process all objects on or near the surface – e.g. hands, individual fingertips, and other objects such as documents. Early examples are Krueger's VideoDesk [16] and Wellner's DigitalDesk [34].

Bottom-up approaches [13, 21, 36] place the camera underneath or behind the surface, and typically employ rear projection onto a diffuse surface material. In most cases, the use of the diffuser attenuates the camera signal, and consequently requires Infrared (IR) techniques [13, 21] to sense IR reflective objects such as fingers on the other side of the surface. This makes sensing arbitrary objects difficult, unless IR visual markers [21] are used to passively identify objects placed on the surface. Conversely, this reduced signal also improves the accuracy of such systems, reducing ambiguities, for example caused by an object far from the surface being accidentally detected by the camera.

This issue of detecting the proximity of objects to the surface is a real challenge for top-down camera systems, particularly in the context of detecting when fingers are touching as opposed to hovering on the surface. In DigitalDesk a microphone was used alongside the camera to coarsely detect when fingers were touching the surface; other techniques (e.g. [19]) often rely on the finger dwelling or other gestures to detect when a user is touching the display.

As demonstrated by Wren et al. [39], TouchLight [36] and the Visual Touchpad [20], stereo vision can assist in detecting the depth of objects in a scene, using various algorithms to compute the disparity between the images captured by the stereo pair. By setting appropriate thresholds it is feasible to detect contact with the surface to within a few centimeters of accuracy.

PlayAnywhere [37] uses a single top down camera and projector set off-axis to detect multi-touch and tangible objects. The system uses an IR camera and IR source to illuminant the foreground objects of interest. The projector plus illuminant guarantees that shadows will be cast on the surface in a variety of lighting conditions. PlayAnywhere uses this shadow information to detect when a single finger from each hand is touching the surface. This can be achieved with millimeter accuracy, and combined with an optical flow technique for bimanual interaction. Tangible objects can also be supported using a fast visual barcoding scheme.

Although the Visual Touchpad is an indirect input device for interacting with large displays in the same space, it shares many commonalities with C-Slate. A stereo camera is employed over a large opaque and darkly colored touchpad to detect touching fingers. Vision techniques are used to track the position and orientation of the hands and make informed guesses as to the identity of each finger to calculate particular gestures. Further UI feedback is provided by segmenting the hands from the images and augmenting them transparently onto the large display. A similar user experience is provided in the TactaPad [28], using optical sensors embedded in the touchpad. A logical progression from Visual Touchpad and TactaPad is to consider how superimposing of hands can be used for feedback in remote collaboration.

2.2 Remote Gesturing Systems

Remote gesturing within CSCW is a growing area of research (see [15] for a detailed review). Early examples are VideoDraw [31] and Clearboard-1 [14], which overlaid analogue video of hands (and in the latter case, upper bodies) of remote participants on a horizontal workstation. Clearboard-2 [14] provided a digital instantiation of the shared workspace, using rear-projection for display, and a camera above the surface. A half mirror and polarizing film were placed above the display surface, which ensured only images of the top half of the participant (and not the screen) would be captured and relayed onto remote parties.

This seminal work combined video conferencing output (focusing on the face) and the surface interaction space (focusing on the hands and arms) into a single display. The system overlaid digital ink on top of this video. In experiments, users found this approach visually overloading at times. Clearly, rendering the video conferencing session underneath the digital content can be distracting. The two channels can interfere with one another, causing difficulties in perceiving the digital content. Further, with this overlaying scheme, occlusion issues exist if the digital content is extended beyond ink, to other types such as documents, images or video. This makes sharing of rich media difficult with Clearboard.

VideoArms [29] combines audio conferencing with a camera pointing at a large surface to capture and segment forearms of people as they interact on screen. These are sent over the network to remote displays where they are overlaid over the shared workspace. The system uses simple computer vision techniques to extract out forearms based on skin tones. Tang et al. describes various revealing experiments with VideoArms where various transparency and rendering effects of virtual arms are evaluated.

Agora [17] and Kirk [15] describe two physical setups using projectors and cameras over a table. Video of physical interactions occurring on a table are captured, and overlaid remotely by projection. Video conferencing facilities are also provided. Kirk [15] and Luff et al. [17] provide quantitative and qualitative evidence as to why such systems can benefit remote collaboration.

Even more recently, the DiamondTouch [8] multi-touch technology has been used for remote collaboration [6, 10]. For example, Digitable [6] combines this multi-touch ability with cameras to provide virtual embodiments of remote users' hands and arms. T3 [32] provides similar mechanisms but uses a camera and projector array with support for input using multiple Anoto pens. These recent systems, share a common goal: to support remote collaboration across horizontal multi-touch surfaces. We aim to explore the combination of this with the power of sensing a variety of tangible objects for both local and remote interactions. Our system shares this motivation with PlayTogether [38], an extension of the PlayAnywhere tabletop system to support remote interactions. Wilson's work serves as an exemplary case that highlights the use of remote gestures and object sharing for gaming.

3. Introducing C-Slate

The C-slate provides fluid techniques for interacting both with the remote collaborator and with the digital workspace. The system is intended to be used in a diverse set of scenarios, from collaborative group work through to gaming. One of the primary goals however is to support collaborative reviewing and annotation of shared electronic documents, for example maps, architectural plans or academic papers. This focus suggests the need for a stylus based interface. For our purposes, we chose to utilize a large commercially available tablet surface, capable of high resolution display and stylus input. This provides new interactive, form factor, and display possibilities when compared to projection.

As shown by [6, 14, 15, 17 29, 31, 32] providing virtual embodiments of arms and hands alongside audio and video conferencing channels can provide much utility for remote collaboration. We therefore attach a camera above the tablet, which faces the surface and captures images of hands and other objects placed over the display. The foreground objects are segmented out from the captured images, and transmitted over the network to remote C-Slates, where they are visually overlaid on the tablet workspace. Unlike some existing remote gesture systems, our approach scales to physical objects as well as forearms, allowing both virtual representations of the users' hands and physical objects to be shared with remote participants. Based on findings of Clearboard, we avoid visually overloading the tablet UI by physically demarcating the video conferencing output from the interaction space using a second display at eye level.

Like PlayTogether, Digitable and T3 we further extend the utility of remote gesturing systems by adding an additional sensing step before transmitting the video for remote rendering. New computer vision and machine learning techniques are applied to each segmented image in order to extract out information regarding recognizable physical object classes. Upon recognition, these objects can be used to carry out local UI actions, which are also relayed onto the remote workspace. This is a key novelty of C-Slate, allowing automatic real-time detection of untagged everyday objects, and allowing online addition of arbitrary new object classes. This recognition system is extended also to distinguish a user's hand from other objects, and further, recognize hand poses, allowing these to also be mapped onto UI actions.

Like [20, 36, 39] we also employ a stereo setup using disparity to estimate when objects are touching the surface. However, for many multi-touch applications, more detailed and high-precision finger touch detection is necessary. The use of a tablet rather than projection makes it difficult to employ a shadow based technique such as in PlayAnywhere. Instead C-Slate provides a new real-time stereo vision and machine learning technique for accurately detecting fingertips and touch. Unlike PlayAnywhere and Visual Touchpad touch is detected for each fingertip, and at a higher level of accuracy than systems that use disparity alone. This allows very fine grained gestures based on individual fingertip data to be integrated into the UI, and additionally allows these gestures to have a remote visual embodiment and effect.

This provides C-Slate surfaces with both multi-touch sensing and high resolution stylus input. Although systems [8, 11] have provided pen and multi-touch input, our configuration provides higher resolution and non-tethered use. This coupled with the ability to detect untagged objects on the surface provides for an exciting new tabletop technology. Additionally, these features are coupled with remote gesturing and object sharing allowing us to explore new remote collaborative scenarios exploring document writing, annotation and editing tasks.

3.1 Shared Workspaces

As shown in Figure 2, the tablet screen provides a window onto a large digital workspace, which is shared across the network with other connected C-Slates. Within the workspace, media items such as images, video, web pages and documents can be opened, reviewed and annotated.

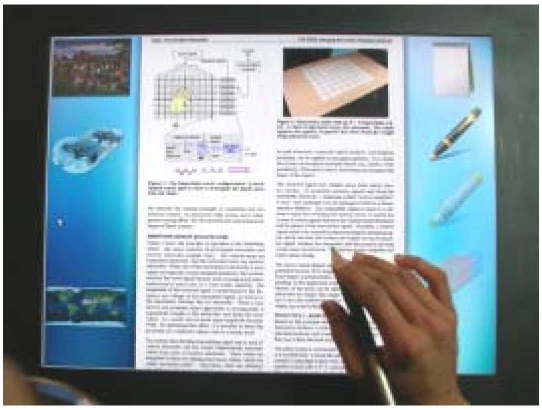

Figure 2: A screenshot of the shared workspace UI. Far left, the media items that can be opened. Far right, controls for switching between pen, highlighter and eraser, and starting a whiteboard session. A selected document is reviewed & annotated in the middle.

By default everything in this digital workspace is shared remotely, i.e. every UI event generated locally is transmitted over the network and rendered remotely. However, the shared workspace is just another application on the user's private desktop, and can be quickly minimized and reactivated to switch between private and public work.

Once connected to other C-Slates, an audio and video conferencing link is also established. Audio in particular has been found to be critical in remote collaboration [15, 17]. Either of these channels can be disabled by the user during collaboration.

3.2 Image Segmentation

Segmenting out the foreground objects in the overhead camera images allows us to selectively overlay content on the remote user's workspace. We make use of a very simple and low-cost optical technique for segmentation, inspired by [14]. This exploits the fact that light emitted by an LCD screen is polarized. We place linear polarization filters on our camera lenses and rotate these so that they suppress the light exiting from the LCD but let other light through (for example from foreground objects). This essentially 'switches off' the display to the cameras, providing a black uniform background which greatly simplifies foreground object extraction.

This approach works more effectively than skin tone analysis used in VideoArms and preserves foreground

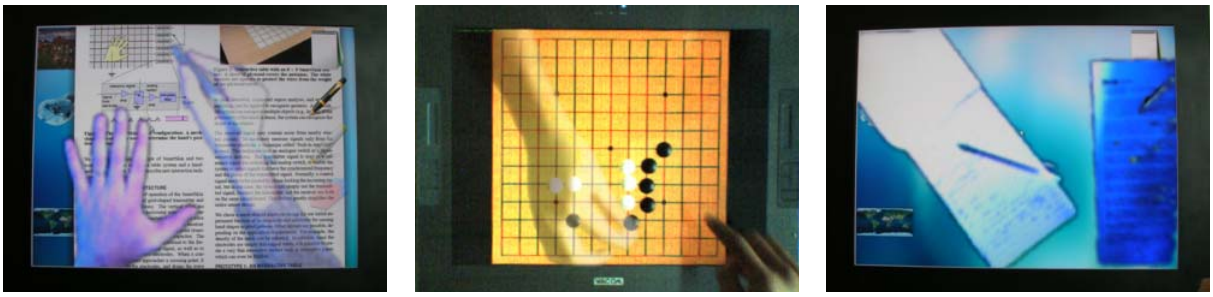

Figure 3: Phantom presence allows each remote user to see the other's hands and objects: (Left) A remote hand holding a pen fading into view on the tablet, whilst the other hand rests on the surface remaining opaque; (Centre) a game of GO being played with a remote participant with a mixture of remote Virtual (white) and local Physical (black) counters. (Right) A real notepad and document placed on the surface are rendered remotely.

color unlike PlayTogether. The technique also scales to segmentation of physical objects placed on the surface.

Before segmentation, the images from the two cameras are first rectified – in our current implementation this is achieved through the camera hardware. Additionally, before rendering objects on the remote screen, the skewed images from the top-down camera are transformed via a homography that is computed beforehand by automatically detecting the corners of the screen in the two images. This allows the overlaid objects to be aligned with the contents of the screen.

3.3 Phantom Presence

The remote gesturing part of C-Slate allows for interactions and objects on the tablet surface to be captured and rendered across remote workspaces. Figure 3 (left & center) shows examples of a remote user's hands being rendered on the shared workspace.

The transparency and blurring of any object (such as hands) is determined by the height of the object from the tablet surface – the closer the object gets to the display the more opaque and sharp it becomes. This gives objects a ghostly effect as they approach the surface, which we call Phantom Presence. This extends the transparency functionality provided in VideoArms, allowing users to get a sense of depth of the remote object. This acts as another peripheral channel for fine-grained coordination whilst also allowing users to mitigate occlusion issues rapidly by controlling the transparency of the object e.g. users may wave their hands high above the surface to gain floor control without completely disturbing the user interacting in the workspace, they may see an approaching hand and coordinate their interactions, or move an object placed on the surface higher up, to allow others to see the content underneath it.

We use stereo vision to calculate the depth of any object. After segmentation of the images from the two cameras, a matching process is used to compute disparity values for each pixel in the foreground. Subtracting the disparity value of the screen at each pixel from this then gives a relative disparity measure that increases directly with height from the screen. These disparity values can be normalized and used to create an alpha mask for the object.

Depending on the application, the users may see exactly the same, or flipped or mirrored versions of the workspace e.g. when used for playing a game of GO, the workspace is flipped on one of the tablets so that the two players see the board from their respective sides. The image of the overlayed hands and objects is also correspondingly flipped to simulate the physical setting where the players would sit on opposite sides of the board.

Phantom presence allows users to share physical objects across the network by rendering images of the artifacts on remote workspaces. Figure 3 (right) shows a document and notepad placed on top of the surface being rendered remotely, thus providing a mechanism for participants to show each other physical artifacts that they are referring to in conversations.

3.4 Object Recognition

Current tabletop systems have demonstrated the utility of using tangible objects for invoking UI actions [21, 23, 37]. Such systems typically require a priori active or passive tagging of objects. We use stereo vision and machine learning to recognize different object classes, by training with labeled images of appropriate untagged objects. Object recognition opens up a new range of possible actions and extends the interactive possibilities of our system. For instance, placing a camera on the surface may enable the pictures to be transferred to disk and shared with the remote participant. Recognizing documents could automatically switch the top-down camera to high-resolution image capture mode intermittently, in order to capture a higher definition image of the document for remote rendering or OCR. Objects could also be used alongside gestures for rapid UI actions, e.g. placing down scissors on the surface and defining a region could cut the text from the document.

The recognition system is adaptive and capable of learning new object classes at run-time. It can also be manually corrected through the UI, to be retrained. The system recognizes objects classes as opposed to particular

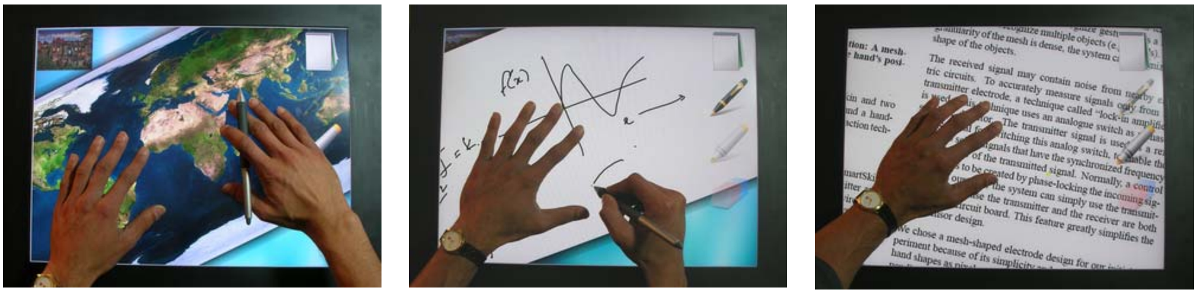

Figure 4: Enabling multi-touch sensitivity using stereo vision. Our detailed analysis of fingertips allows for symmetric and asymmetric bimanual interaction: scaling, rotation, translation with simultaneous use of the stylus and also single handed rotation and scaling.

instances of these objects – this has particular tradeoffs as discussed later.

For the automatic classification of an image into one of several possible objects, we use a dataset of pre-labeled images of the various object classes that we support (up to 24 classes at present). A model based on random forests [4] is learnt in order to optimally discriminate between the different objects. This involves automatically learning a set of tests (rules) in the form of a number of decision trees [7]. Having segmented the image, each foreground connected component is taken as a separate object. At test time each foreground pixel undergoes the tests encoded in each of the learnt trees and a histogram distribution over leaf labels given by the random forest is computed for each object. The final classification step is performed by nearest neighbor comparison of each such histogram with exemplar (training) histograms using a distance metric which is invariant to rotation or mirroring of the object.

Our classification algorithm combines appearance, shape and depth cues to achieve accurate class discrimination in real-time. In particular, stereo features (expensive in general) are computed on an on-demand basis, only if necessary for discrimination. Our current automatic algorithm achieves recognition accuracy around the 99% mark on 24 object classes. More details are provided in [7].

A coarse level touch-detection is also obtained for these objects using the disparity cues. The classifier can also be trained to identify hands and particular hand poses. For example, hand poses being could be used for actions such as virtual object selection by pointing and closing a window using a fist. These events are only generated when the user's hand is touching the surface (computed from depth information from the stereo camera). The system is also capable of recognizing the hand holding an object. For example, the system can recognize an eraser and highlighter in the hand, and allows simple interactions to be mapped to these e.g. highlighting or erasing the text underneath the hand when the object makes contact with the surface.

3.5 Gesture and Bimanual Interaction

The pose recognition does not work at a precision level of individual fingertips, and although this is suitable for many simple UI actions, more fine-grained analysis is required to support other common multi-touch gestures [3, 40]. We again use the stereo camera to enable multi-touch on the tablet, coupling this with simultaneous high resolution stylus input. This allows us to support asymmetric interactions [12] using the stylus in the dominant hand whilst carrying out peripheral actions with the non-dominant hand [11, 12], as well as symmetric interactions more classically associated with bimanual interfaces [3].

Figure 4 shows some of the potential interactions: two handed symmetric scaling and rotation (note the stylus, though present, is not used in this case); asymmetric action involving rotation with the non-dominant hand while using the stylus with the dominant hand; and finally zooming into a document with a single hand.

Once the system recognizes a hand, a more detailed analysis is carried out to detect individual fingertips and accurately determine if they are touching the surface. Visible finger tips are detected in the image via a two step process. Individual points on the edge of the segmented hand are first classified as lying on a finger tip or not, and then a clustering phase combines evidence from multiple points to detect the finger tips. The classifier is built again by using a training database of images of hands, labeled with the fingertip positions. Each point is robustly encoded as a 64 dimensional signature vector computed from a local image patch around it and a Support Vector Machine [22] is trained to distinguish between tip and non-tip points based on these signatures.

Detected fingertips are further processed to detect touch. In our setup, conventional stereo algorithms that compute disparity images (e.g. [20, 39]) fail to meet the level of precision for our requirements – we found often when a finger hovers above the screen it can be falsely detected triggering unexpected behavior in the UI. We have developed an algorithm that probabilistically aggregates stereo cues from several points at each fingertip and uses a finger-specific geometric model to resolve this. The method returns a probability value for the touch detection of each finger and allows multi-touch sensing with several millimeter precision. Details of the approach are described in full in [1].

4. Implementation

The C-Slate is realized using off-the-shelf hardware, although the frame attaching the various components together has been custom built. Wacom's Cintiq 21UX is used for the tablet display, providing a large 21" display size and 1600x1200 screen resolution with a digitizer capable of detecting strokes, hover and pressure from a single stylus. We use the Bumblebee 2, a stereo camera from Point Grey Systems, capable of providing high framerate, hardware synchronized and rectified stereo images. We have also tested our algorithms with a pair of standard web-cameras to form a stereo camera. These are found to work satisfactorily although they provide slightly lower framerates and accuracy. Our applications are built using a combination of Win32 and WPF. Phantom Presence video is streamed as encoded low-res PNG images over UDP preserving per-pixel alpha values. UI updates, such as mode switches, ink strokes, or changes in the transform matrix of a virtual object, are sent between shared workspaces using a proprietary protocol again using raw sockets.

5. Discussion and Future Work

In this paper, we have focused on describing the underlying technology of C-Slate. As demonstrated we have used new stereo vision techniques in various ways including remote gesturing, touch detection, and object recognition to create a new tabletop technology that also supports remote collaboration. There are of course interesting issues and observations associated with the use of our computer vision based systems. Adverse lighting is clearly an issue for vision systems. This is in part mitigated by attaching a light source to the C-Slate stand, but other approaches are also being explored. Occlusion is also an often cited problem of using overhead cameras on interactive surfaces. However we have found it to be far less of an issue than we originally anticipated. Perhaps telling is that the majority of the users so far have thought it the tablet is in fact a touchscreen rather than a camera-based system. We partly attribute this to the fine-grained touch detection enabled on the surface, but would attempt to quantify this in our future studies.

System robustness could be further improved for touch detection by supporting automatic online addition of new training examples. For example, we could couple our camera-based technique with a single-point touch overlay on our tablet screen, which allows us to unambiguously determine if a user is touching the surface and remove any false positives from our classifier.

Sharing of paper documents appears to be an important aspect of remote collaboration. Our cameras do not have the resolution to sufficiently capture detailed images of documents but we anticipate integrating a high resolution stills camera to the setup to improve image clarity. Automatic detection of a paper document could programmatically trigger the camera to take an image. This high resolution image could be incorporated into the Phantom Presence video in a lightweight manner by transmitting the high resolution image once, and using paper tracking techniques [37] to render this onto the outline of the object.

For a full assessment of our system, we are planning to run extensive user studies. In the interim, we have begun deploying the system in our workplace to get initial feedback. Although not quantified, people have expressed how C-Slate offers a natural and expressive way of collaborating remotely. For example, at a glance users can see if a remote participant is interacting on the screen, writing, pointing and referring to something, waiting for a response and so forth. The transitions between these states are also easy to perceive e.g. switching from writing to pointing. The visual impact of seeing the hand makes it easier to draw peoples' attention and less prone to being lost on visually noisy backgrounds. The transparency effect allows users to gain more peripheral awareness of when peoples' hands are approaching the surface. This appears to make coordination tasks much easier.

Users have also been enthusiastic about the bimanual interaction techniques. The asymmetric and symmetric bimanual input enabled by simultaneous touch and stylus input indeed opens up another channel for interaction – the non-dominant hand is frequently used for both local and remote interactions. It is very intuitive for a user to be writing with one hand and gesturing with the other. Rotation and translation with non-dominant hand seems particularly important during writing tasks while scaling seems secondary and sometimes distracting during writing.

The ability to detect object classes presents a novel and potentially powerful feature for users, allowing them to quickly invoke UI actions using physical objects that already have strong meanings and affordances associated with them. E.g. a user is likely to readily understand the actions to expect when bringing a physical highlighter or eraser to the digital surface. This however indicates the need for careful UI design, as physical objects could potentially cause unanticipated UI behavior.

Finally we are investigating other interactive arrangements for remote group work, for instance where multiple co-located participants are interacting with remote parties using C-Slate. These will require new tabletop configurations, for example tiling two tablet displays together to increase the size of the physical workspace for group interaction.

6. Conclusions

We have presented a novel set of technologies designed to improve remote collaboration on horizontal surfaces.

Initial user feedback on the system has been positive and extensive user studies are on our agenda. The contributions of this paper are a new multi-touch and object sensing system called C-Slate that supports multimodal input from untagged objects, high precision multi-touch, hand poses, and high-resolution stylus input on a large tablet surface, thus providing a new interactive and interesting tabletop technology; and the exploration of these tabletop systems for remote collaboration combining and extending the work carried out by the tabletop, multi-touch, and remote gestures communities.

Acknowledgements

We thank Andy Wilson, Abigail Sellen, Bill Buxton and Dave Kirk for all their inspiration and feedback.

7. References

1. Agarwal, A. et al. High Precision Multi-touch Sensing on Surfaces using Overhead Cameras, In Tabletop'07.
2. Apple iPhone Multi-touch, http://www.apple.com/iphone
3. R. Balakrishnan and K. Hinckley. Symmetric bimanual interaction. In ACM CHI, pages 33–40, 2002.
4. L. Brieman. Random Forests. Machine Learning, 2001.
5. W. Buxton. Multi-Touch Systems that I Have Known and Loved. http://www.billbuxton.com/multitouch
6. Coldefy, F.; Louis-dit-Picard, S., DigiTable: an interactive multiuser table for collocated and remote collaboration enabling remote gesture visualization, In ProCams '07.
7. T. Deselaers et al. Incorporating On-demand Stereo for Real Time Recognition. In CVPR, 2007.
8. Paul Dietz and Darren Leigh. DiamondTouch: a multi-user touch technology. 2001.
9. P.H. Dietz et al. DT Controls: Adding Identity to Physical Interfaces. In Proceedings of UIST, 2005.
10. Esenther, A., and Ryall, K., RemoteDT: Support for Multi-Site Table Collaboration. In CollabTech 2006.
11. Y. I. Gingold, et al. A Direct Texture Placement and Editing Interface. In UIST 2006.
12. Y. Guiard. Asymmetric division of labor in human skilled bimanual action: The kinetic chain as a model. The Journal of Motor Behavior, 19(4):486–517, 1987.
13. J. Y. Han. Low-Cost Multi-Touch Sensing through Frustrated Total Internal Reflection. In UIST, 2005.
14. H. Ishii, M. Kobayashi, and J. Grudin. Integration of Interpersonal Space and Shared Workspace: ClearBoard Design and Experiments. ACM TOIS, 11(4), Oct 1993.
15. David Kirk. Turn It This Way: A Human Factors Treatise on the Design and Use of Remote Gestural Simulacra. PhD Thesis, University of Nottingham.
16. Krueger, M, Videoplace: an artificial reality. In CHI 1985, pages 35–40.
17. H. Kuzuoka, J. Kosaka, K. Yamazaki, Y. Suga, A. Yamazak, P. Luff, and C. Heath. Mediating Dual Ecologies. In Proceedings of CSCW 2004.
18. JazzMutant Lemur. http://www.jazzmutant.com/lemur_overview.php.
19. J. Letessier and F. Berard. Visual Tracking of Bare Fingers for Interactive Surfaces. In UIST 2004.
20. Shahzad Malik and Joe Laszlo. Visual Touch-pad: A Two-handed Gestural Input Device. In ICMI 2004.
21. Microsoft Surface, http://www.surface.com
22. John Platt. Probabilities for Support Vector Machines. Advances in Margin Classifiers, pages 61–74, 1999.
23. J. Rekimoto. SmartSkin: an infrastructure for freehand manipulation on interactive surfaces. In CHI 2002.
24. Y. Rogers and S. Lindley. Collaborating around vertical and horizontal displays:which way is best? In Interacting With Computers, pages 1133–1152, 2004.
25. K. Ryall et al. Exploring the Effects of Group Size and Table Size on Interactions with Tabletop Shared-Display Groupware. In CSCW 2004, pages 284–293.
26. S. D. Scott, K. D. Grant, and R. L. Mandryk. System Guidelines for Co-located, Collaborative Work on a Tabletop Display. In Proceedings of ECSCW, 2003.
27. Tactex Controls Inc. Array Sensors. http://www.tactex.com/products_array.php.
28. TactaPad. http://www.tactiva.com/tactapad.html.
29. Tang, A., Neustaedter, C. and Greenberg, S. VideoArms: Embodiments for Mixed Presence Groupware. In BCS-HCI 2006.
30. Tang, John C. Findings from Observational Studies of Collaborative Work. International Journal of Man-Machine Studies, 34(2):143–160, 1991.
31. Tang, John C. and Scott L. Minneman. VideoDraw: A Video Interface for Collaborative Drawing. In Proceedings of CHI, pages 313–320, 1990.
32. Tuddenham, P. Distributed tabletops: territoriality and orientation in distributed collaboration. In CHI '07 Extended Abstracts.
33. V. Vapnik. The Nature of Statistical Learning Theory. Springer, 1995.
34. Pierre Wellner. Interacting with Paper on the Digital Desk. Communications of the ACM, 36(7):86–96, 1993.
35. Wayne Westerman. Hand Tracking, Finger Identification and Chordic Manipulation on a Multi-Touch Surface. In PhD dissertation, University of Delaware, 1999.
36. Andrew D. Wilson, 2004. TouchLight: An Imaging Touch Screen. In ICMI 2004.
37. Andrew D. Wilson. PlayAnywhere: A Compact Interactive Tabletop Projection-Vision System. In Proceedings of UIST, 2005.
38. Andrew D. Wilson and D. Robbins. PlayTogether: Playing Games across Multiple Interactive Tabletops. 2007.
39. Wren, C. and Ivanov Y, Volumetric Operations with Surface Margins, In CVPR 2004.
40. Wu, M. et al. Multi-finger and whole hand gestural techniques for tabletop displays. In UIST 2003.